Assessing bias and uncertainty in the HadAT-adjusted radiosonde climate record

Assessing Bias and Uncertainty in the HadAT-Adjusted Radiosonde Climate Record

MARK P. MCCARTHY, H. A. TITCHNER, AND P. W. THORNE

Met Office Hadley Centre, Exeter, United Kingdom

S. F. B. TETT

Met Office Hadley Centre, Reading, United Kingdom

L. HAIMBERGER

Department of Meteorology and Geophysics, University of Vienna, Vienna, Austria

D. E. PARKER

Met Office Hadley Centre, Exeter, United Kingdom

(Manuscript received 18 October 2006, in final form 27 June 2007)

ABSTRACT

Uncertainties in observed records of atmospheric temperature aloft remain poorly quantified. This has resulted in considerable controversy regarding signals of climate change over recent decades from temperature records of radiosondes and satellites. This work revisits the problems associated with the removal of inhomogeneities from the historical radiosonde temperature records, and provides a method for quantifying uncertainty in an adjusted radiosonde climate record due to the subjective choices made during the data homogenization.

This paper presents an automated homogenization method designed to replicate the decisions made by manual judgment in the generation of an earlier radiosonde dataset [i.e., the Hadley Centre radiosonde temperature dataset (HadAT)]. A number of validation experiments have been conducted to test the system performance and impact on linear trends.

Using climate model data to simulate biased radiosonde data, the authors show that limitations in the homogenization method are sufficiently large to explain much of the tropical trend discrepancy between HadAT and estimates from satellite platforms and climate models. This situation arises from the combination of systematic (unknown magnitude) and random uncertainties (of order 0.05 K decade^{-1}) in the radiosonde data. Previous assessment of trends and uncertainty in HadAT is likely to have underestimated the systematic bias in tropical mean temperature trends. This objective assessment of radiosonde homogenization supports the conclusions of the synthesis report of the U.S. Climate Change Science Program (CCSP), and associated research, regarding potential bias in tropospheric temperature records from radiosondes.

1. Introduction

The most recently produced climate quality homogenized radiosonde datasets (Thorne et al. 2005a; Free et al. 2005; Haimberger 2007) indicate warming throughout the troposphere since 1958. Globally, this is at a similar rate to that reported at the surface (Karl et al. 2006). However, when considering the more recent satellite era (1979 onward), the same radiosonde datasets indicate that the troposphere is warming at a slower rate than at the surface, particularly within the tropics. This is at odds with climate models, which predict amplification of the surface trends in the tropics (Santer et al. 2005; Karl et al. 2006). Some Microwave Sounding Unit (MSU) satellite records indicate similar changes in the troposphere to radiosonde records (Christy and Norris 2006), whereas others are in broad agreement

Corresponding author address: Mark P. McCarthy, Met Office, Hadley Centre, Fitzroy Rd., Exeter, Devon EX1 3PB, United Kingdom.
E-mail: mark.mccarthy@metoffice.gov.uk

DOI: 10.1175/2007JCLI1733.1

with the model predictions (Mears and Wentz 2005; Vinnikov et al. 2006).

Differences between observational estimates of temperature trends in the upper air reflect the difficulty in adequately identifying and correcting for the many undocumented changes that exist (Free and Seidel 2005) and highlight the importance of structural uncertainty arising from methodological considerations (Thorne et al. 2005b). Techniques used to create radiosonde temperature datasets, with the exception of Haimberger (2007), have tended to be manually intensive and use different station selections along with expert judgment and incomplete metadata records. They have the advantage of using considered value judgments based upon all available evidence. Their major limitation relative to other methods is that they require considerable subjective judgment, and are therefore not fully reproducible.

Here we present an automated method for creating radiosonde temperature time series. The system uses a neighbor-based iterative approach similar to the manual method employed to create the current Hadley Centre radiosonde temperature dataset (HadAT; Thorne et al. 2005a). Its purpose is to assess bias and uncertainty in a HadAT-like adjusted radiosonde climate record, and ultimately in estimates of decadal trends, to complement the existing "best guess" bias-corrected datasets. Our system allows for the generation of a large number of possible realizations of the climate data record using a range of methodological assumptions. In this paper we use both radiosonde data and simulated data from a global climate model to objectively assess the effectiveness of the system, and we quantify the main sensitivities of the system and systematic biases that may explain at least part of the apparent surface–troposphere temperature trend discrepancy. To this end, this paper focuses on temperature trends observed by radiosondes in the lower troposphere during the satellite era, although other periods and levels are also included.

2. Data sources

a. Radiosonde

Radiosonde data were collated from a number of sources in the generation of the HadAT set in order to provide as complete data coverage as possible, and to provide a reference network of higher-quality station records to use as neighbor stations in the correction of the more comprehensive network. A full discussion on the input data and their use can be found in Thorne et al. (2005a). In this study we use the ungridded HadAT station time series. HadAT0 comprises uncorrected seasonal mean data from 476[1] stations of mixed sources and observation times. HadAT1 comprises the same 476 stations after homogenization.

In addition to HadAT we have used data and metadata from the Integrated Global Radiosonde Archive (IGRA; Durre et al. 2006). Station soundings for the period 1958 to 2003 were used. Monthly means were computed for daily 0000 and 1200 UTC launches at 14 pressure levels (1000, 850, 700, 500, 400, 300, 250, 200, 150, 100, 70, 50, 30, and 20 hPa) where at least eight ascents were recorded in a given month. A biweight mean (Lanzante 1996) was used to reduce the influence of outliers. Seasonal means were calculated where at least two out of three monthly means were available, and each station series was converted into anomalies with reference to a 1966–95 climatology period. We excluded stations and levels that did not have at least five years of data containing at least three seasons for each of the three decades within the climatology period. We also excluded Indian stations because they have been found to be difficult to homogenize (Thorne et al. 2005a) and problematic for subsequent analysis of long-term trends (Parker et al. 1997; Lanzante et al. 2003). The 0000 and 1200 UTC soundings were combined to produce a comprehensive merged dataset of 509 stations, 50 of which are within the (20°S–20°N) tropics.

A main advantage of IGRA over HadAT is the ability to separate the daytime and nighttime data. Therefore, day and night datasets were generated using a simple criterion that 90°E–90°W is daytime for 1200 UTC and nighttime for 0000 UTC, and vice versa for all other longitudes. The stations were limited to between 70°N and 70°S to avoid the seasonality of polar day and night. This gave a total of 465 stations for the daytime and 384 stations for the nighttime IGRA datasets.

b. Simulated data

To investigate the capabilities of the homogenization system, we have used data from the Third Hadley Centre Atmospheric Model (HadAM3: Pope et al. 2000). The model data were forced with observed sea surface temperature and sea ice distributions from the Hadley Centre Sea Ice and SST (HadISST) dataset (Rayner et al. 2003) for the period 1978–99. In addition, the model included forcings from changing solar output, stratospheric aerosols from volcanic eruptions, tropospheric and stratospheric ozone, greenhouse gases, land surface, and sulfate aerosols (Tett et al. 2007).

[1] The HadAT literature refers to 477 stations. However, a duplicate station was found and removed.

For each of the 476 HadAT station locations we created seasonal mean anomaly time series from the model grid box within which each station falls. Anomalies were taken with respect to the entire model period. Model data were available for eight pressure levels (850, 700, 500, 300, 200, 150, 100, and 50 hPa). We added random noise with a Gaussian distribution and standard deviation half that of the model grid box. This was done to ensure that simulated series from stations that fell within the same model grid box were not identical, but were still highly correlated. The resulting model dataset has the same spatial and monthly-mean sampling as HadAT, but contains no instrumental break points. The average of the ratio of standard deviations for individual station temperature time series from the model data compared to radiosondes is 1.09. This lends support for the use of the model data as a surrogate for observations in our validation exercise.

3. Method

HadAT was generated using an iterative neighbor-based homogenization. Break points were identified by manual analysis of statistical and metadata evidence of spurious step changes in the data record. We have automated this process so that we can objectively test the sensitivity to several methodological assumptions, which we now describe.

The basis for this, and many other breakpoint detection schemes, is to test the null hypothesis, H_0, that there are no break points. For this purpose a break point is defined as a change in the mean value of the time series that is a direct result of a change in instrumentation or observing practice. To test the null hypothesis we use two pieces of information, the probability of rejecting the null hypothsis from a statistical breakpoint identification, S, and a probability from the metadata record of known changes at a given station, M. We define the joint probability of obtaining M and S given the null hypothesis as

$$P(M \cap S|H_0) = P(S|H_0)P(M|H_o). \quad (1)$$

Equation (1) assumes that M and S are dependent only through the break points (i.e., they are conditionally independent). The joint probability [left-hand side of Eq. (1)] at each point in the time series is then used to indicate the points where the null hypothesis may be rejected.

For this study, as in HadAT, we use a nonparametric Kolomogorov–Smirnov (K–S) test (Press et al. 1992) as a statistical homogeneity test. This is applied to time series of seasonal mean differences between station data and weighted composites of data from neighboring stations. The weighting of each contributing neighbor station is equal to the correlation coefficient between seasonal mean anomalies from reanalysis fields [National Centers for Environmental Prediction–National Center for Atmsopheric Research (NCEP–NCAR; Kalnay et al. 1996) or the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40; Uppala et al. 2005)] sampled at station locations over the period 1979–2003. Neighbors are selected only from locations that fall within a contiguous area with correlation coefficients greater than $1/e$ surrounding the target station. The neighbor weightings are calculated in the same way as those used in HadAT; for further details on their determination and limitations see Thorne et al. (2005a). A key assumption here is that the neighbor reference series is a reasonable estimate of the common natural variability between the target station and its neighbors. This ideally requires that break points in contributing neighbor stations are randomly distributed about zero in value and occur randomly in time so that their impact is minimized through the process of averaging.

Previous studies (Gaffen et al. 2000) concluded that neighbor-based checks were inappropriate for radiosondes due to the large station separation and coincident break points within countries. We tested the suitability of using near-neighbor reference series by running our system on the radiosonde station series in isolation rather than using station minus neighbor differences. In the station-only case it was found that trends were completely removed from all levels. In the neighbor-based system large-scale trends were increased and decreased for different regions and time periods, and the vertical trend profile was grossly retained. Therefore, the use of a neighbor-based reference series for the detection and correction of break points is a sufficient constraint on the large-scale mean trends, but the limitations are further discussed in section 5.

There exist multiple K–S test statistics, one for each pressure level, $P(L_1)$, $P(L_2), \ldots, P(L_n)$. To maintain consistency between the individual pressure levels, the $P(S|H_0)$ component of the breakpoint detection algorithm is estimated from the geometric mean of the available K–S statistics:

$$P(S|H_0) \propto \left(\prod_{k=1}^{n} P(L_k) \right)^{1/n}. \quad (2)$$

We use a simple subjective probability model to estimate $P(M|H_0)$. We set a background value of 1, with each metadata event represented as an inverted Gaussian curve (see appendix A for alternatives) with a cutoff

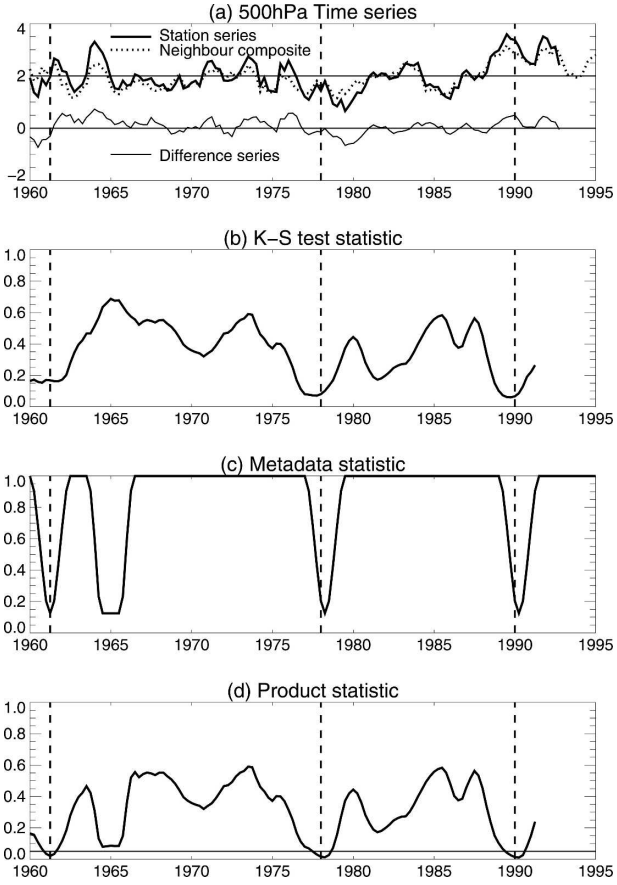

FIG. 1. The automatic breakpoint identification method for station 03774 (Crawley, UK). (a) The original temperature data at one level only (500 hPa). The station (thick solid), neighbor composite (dotted), and station minus neighbor difference (thin solid) are shown. The station and neighbor series are offset by 2 K from the difference series, for clarity. (b) The geometric average of the K–S statistic from all nine levels. (c) The metadata probability function (see section 3b). (d) The product of (b) and (c). The horizontal line denotes the critical threshold for detection in this example (see text); dashed vertical lines highlight where break points have been identified.

point six seasons either side of the reported timing of the event. This model therefore accounts for some uncertainty in the reported date and is similar to the model used by Haimberger (2007). Metadata events without specified dates are not used.

An example is shown in Fig. 1. Break points are identified from the product (Fig. 1d) of the K–S statistic (Fig. 1b) and metadata statistic (Fig. 1c). We will refer to this as the breakpoint score since it is not strictly speaking a probability. The lower the score the greater confidence we have that a potential break point exists at that time point. We locate periods with a score less than or equal to a predefined critical value for at least three consecutive points so that the break point is a robust feature of the time series and not a numerical artifact of the statistical breakpoint detection. The minimum within each of these periods below the critical value is assigned as the break point. Only one break point is allowed to occur within a predefined period of time (default of 2 yr, but see appendix A for a range of possible values). Figure 1 represents a single step in a system that simultaneously corrects neighbor stations and is conducted iteratively. Therefore Fig. 1a should not be interpreted as an indication of the complete set of break points detected (or not) at this station.

The selection of an appropriate critical value is an important consideration. From an analysis of climate model data, which are highly correlated in the vertical and exhibit autocorrelation, we found that in the absence of break points approximately 5% of seasons in eight level, upper-air temperature data will produce a K–S statistic value of 0.1 or lower. Therefore, the critical value should not greatly exceed this threshold if we wish to minimize the false detection rate.

We estimate the adjustment factor from the time series of station minus neighbor differences. It is calculated as the difference in the medians of predefined periods (default 10 yr, also see appendix A) either side of an identified break point. Should another break point exist within this adjustment window, then the period is reduced so as not to span any other break points. At least five seasons of nonmissing data are required either side of the break point in order for an adjustment estimate to be calculated, otherwise it is ignored.

A critical step in the generation of HadAT was the manual inspection of the statistical and metadata evidence in the application of adjustments. This was required to confirm that the adjustment estimates were well defined and not influenced by break points in the neighbor composite or outliers in the neighbor or station time series. We have attempted to replicate this decision-making process as closely as possible with a number of simple tests also used in the generation of HadAT (Thorne et al. 2005a). These tests are described in appendix B. If the break point fails this set of tests, then the adjustment is not applied.

The system is run iteratively. The adjusted data from each iteration are fed back through the system, recalculating neighbor composites each time. The neighbor composites should therefore improve as subsequent iterations are conducted. In early iterations we set a very low critical value threshold for break points so that only the worst break points are identified. In later iterations, after these worst offenders have been removed, we relax this threshold to detect smaller break points, or recalculate adjustments that were rejected in earlier iterations that are now better constrained.

The automated system is critically reliant on a number of parameters that will directly or indirectly influence the number of break points detected, false detection rates, and adjustment estimates. These parameters are summarized in appendix A. In the first instance we set them to values that most closely resemble those used in the manual generation of HadAT (sections 4 and 5). We then investigated the impact of changing these parameters within reasonable bounds to investigate methodological uncertainties and how these impact the recovery of large-scale trends (sections 6 and 7).

4. Comparison with HadAT

To be able to infer conclusions about uncertainties in HadAT (or similarly constructed datasets) we need to show that, given appropriate tuning, the automated system can reasonably replicate this manual process. Therefore, we passed the unadjusted HadAT0 data through our system, and compared the properties of the breakpoint detection, adjustments, and large-scale mean trends with those in the adjusted HadAT1.

The results of this are summarized in Fig. 2. A positive detection of a break point was defined as one that occurs within one year of a HadAT break point. Sixty-one percent of all HadAT1 break points were detected, with 70% of the break points larger than 0.5 K detected. Figure 2b shows that 14% of the break points detected by the automated system were not found in HadAT, that is, 86% of breaks found by the automated system were found in HadAT. The agreement with HadAT rises to 94% if we relax the time-match criterion to 2 yr and decreases to 75% for a two seasons criterion. The total number of break points found by the automated system (1972) was less than in HadAT (3063). In Figs. 2c and 2d we show that the automated system and HadAT are in good agreement in terms of the adjustment magnitudes, with no significant bias, and a root-mean-square difference of 0.39 K.

In the tropics we find that the impact on trends of both the automated and manual adjustments are similar (Fig. 3) and act to slightly cool the trends relative to the unadjusted data for the period 1958–2003. For global data the differences between the three datasets (unadjusted, HadAT, and automated system) are not statistically significant.

In summary, we have found that the automated decision process, when presented with equivalent evidence, is able to replicate many of the decisions on breakpoint location and magnitude made in the generation of HadAT. This gives us confidence that we can use the automated system to objectively investigate uncertainty inherent in the HadAT method of dataset construction.

5. Validation and system limitations

We have applied a number of simple breakpoint models to the HadAM3 model data to investigate how well the homogenization might be expected to perform under simple assumptions regarding the properties of break points. A total of nine experiments were conducted and the properties of these are given in Table 1.

a. Breakpoint detection

The breakpoint detection statistics are summarized in Table 2. A total of 462 break points were identified in the model data free of spurious break points (UNADJ). These false break points were not entirely randomly distributed through time. The system identified a small number of break points in the vicinity of major climatic events associated with volcanic activity (1983, 1991) and ENSO events. At any single event some 5% of stations were affected. The assumption that neighbor composites capture local, natural climate variability is deficient at some times and locations. This is to be expected given the large distances between some stations. Time-invariant neighbor coefficients are not always appropriate in the presence of intermittent large-scale phenomena, which impose time-varying geographical coherence. However, the small proportion of stations affected means that this is not a major concern for large-scale mean diagnostics and linear trends (see also Fig. 4, and associated discussion in section 5c).

The probability of detection shown in Table 2 is encouragingly high, with experiments RNDM, SKEW, and SMALLSKEW all detecting more than 80% of break points. The detection is higher if we only consider the larger break points. For example RNDM captures 98% of breaks >0.4 K. The loss of metadata in META results in a drop in detection rates. This primarily affects the smaller break points. Increasing the density of break points in experiments MULTI, MULTISKEW, and HATA reduces the breakpoint detection for all break sizes because the neighbor composite reference series are more likely to be contaminated. In MULTISKEW 95% of breaks are <0.4 K, which also results in a significantly lower detection rate. The false detection rate is between 24% and 34% for most experiments. HATA and META have the highest false detection rates, indicating that metadata are important in both positive and false detection rates.

There is an additional concern relating to the vertical coherency of break points. Break points can, and do, occur at individual levels (Lanzante et al. 2003; Thorne et al. 2005a). To test this we reran the RNDM experiment, but only applied the break points to the 500-hPa level. For this single-level breakpoint experiment the

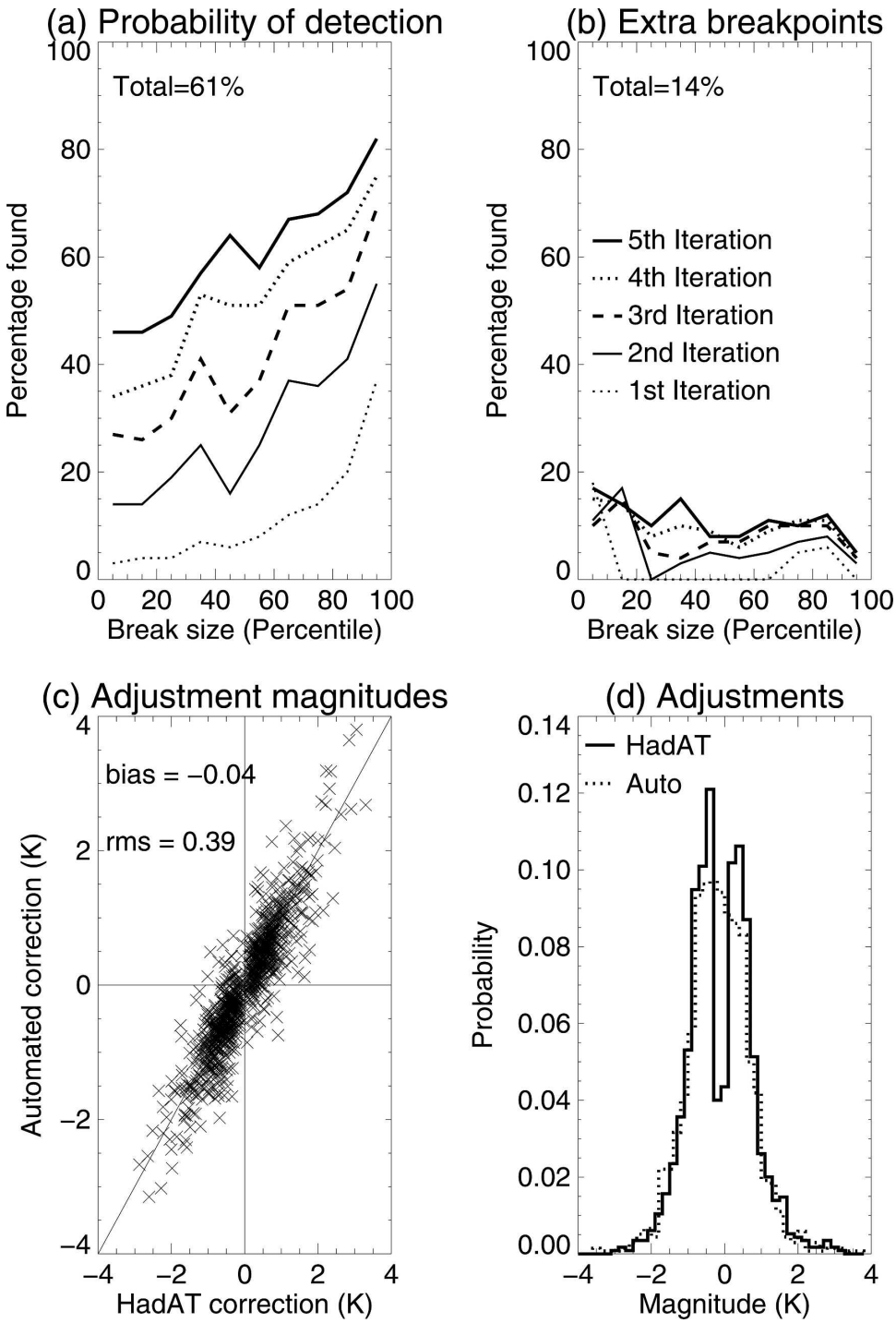

FIG. 2. (a) Probability of detection of a break point within one year of a HadAT1 break point from our automated system at each iteration. The probability is estimated as the ratio of HadAT1 break points found by the automated system to all HadAT1 break points. The total detection rate is also given. (b) Probability of detection by the automated system, but not by HadAT. The probabilities of detection are shown against the percentile of the breakpoint magnitude. The 10th and 90th percentiles are 0.19 and 1.2 K, respectively. (c) A scatterplot of the coincident HadAT1 and automated break points in terms of the estimated adjustment magnitude. The mean bias and RMS error between the automated and HadAT1 adjustments are shown. (d) The normalized histograms of adjustments made by HadAT1 (solid) and the automated system (dotted).

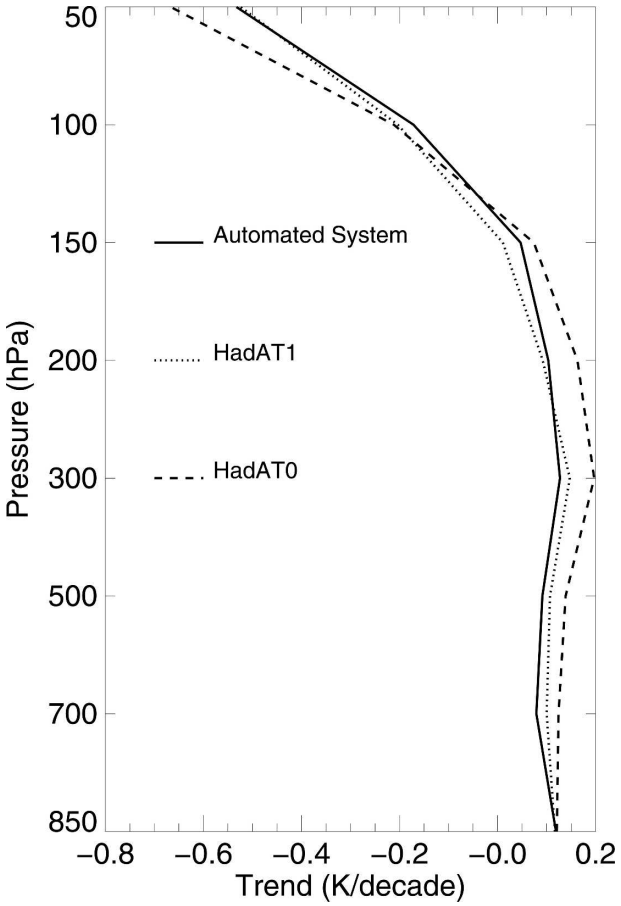

FIG. 3. Profile of trends from (dashed line) HadAT0, (dotted line) HadAT1, and (solid line) the automated homogenization system. Trends are median of pairwise slopes trends for the tropical mean temperature series at eight atmospheric pressure levels from 1958 to 2003.

probability of detection was 59%, with a false detection rate of 34%. Single-level break points are harder to detect with our system, but they are still detected in most cases, so we do not consider this to be a major flaw in our methodology.

b. Adjustment

The uncertainty in the estimation of the adjustment is largely independent of the magnitude of the break point (e.g., Fig. 2c). In other words, the likely magnitude of error for a 2-K break point will be the same as that for a 0.2-K break point. However, it is sensitive to the distribution of break points in both space and time within the network of stations, as shown in Table 2. We expect the largest contribution to random error to come from the inadequacies in the neighbor reference series due to natural climate variations (UNADJ), coincident break points at neighbor stations (CNTY), or a high density of break points (MULTI), and the values in Table 2 support this. The average root-mean-square-error of all experiments is 0.4 K, suggesting that we cannot resolve break points to better than 0.4 K, a value also supported by the observations in the HadAT comparison (Fig. 2c).

For most of the experiments there is little or no bias (the average difference between the estimated and applied adjustments) in the adjustments, suggesting that any residual inhomogeneities should be random regardless of the distribution of the actual break points. However, in the presence of a pervasive bias within the dataset, as in SKEW and MULTISKEW, there is a systematic underestimate of breakpoint magnitudes of up to 0.1 K, which will result in a sign bias in the residual errors that remain following the homogenization. This has important implications for the recovery of trends in biased datasets using the HadAT-like method of data homogenization.

c. Trends

Both the model and observations have a vertical trend gradient, with strong stratospheric cooling (-1 K decade^{-1} at 50 hPa) and weak tropospheric warming ($+0.15$ K decade^{-1} at 300 hPa), but we find that for these tests, where the break points are identical at all levels, the impact of the homogenization is largely independent of pressure level. Therefore, our results are summarized as averages across all the model pressure levels and can be considered indicative of any individual level.

The absolute trend error averaged across all pressure levels, before and after homogenization, for each of the nine model experiments is shown in Fig. 4. Uncertainties of order 0.4 K in each adjustment applied to the data, and the presence of false break points results in residual trend uncertainties of 0.02–0.05 K decade^{-1} in experiments without a significant systematic trend bias (UNADJ, RNDM, CNTY, and META). The upper bound in this range is the CNTY experiment, resulting from larger uncertainties in the adjustments made for this experiment (Table 2). These trend uncertainties are similar in magnitude to those presented in Thorne et al. (2005a), which were estimated using a Monte Carlo method of aggregating uncertainties from individual adjustments.

In SKEW and SMALLSKEW the trend bias is significantly reduced by the homogenization process. In the case of SKEW this was achieved by reducing tropospheric trends that were biased positive, and for SMALLSKEW trends were biased negative and therefore increased following homogenization. This confirms

TABLE 1. Summary of the experiment details for assessing the performance of homogenization under a number of idealized conditions.

Expt	Description
UNADJ	The model data were passed through the system without any break points added. Therefore, any break points found were false detections.
RNDM	Each station was given two artificial break points randomly located in time, but separated by at least five years and of random magnitude. The magnitude was from a normal distribution about zero with standard deviation of 0.7 K (estimated from the distribution of HadAT break points). Each break point was applied to all pressure levels, and the system was provided with a metadata record of the location, but not magnitude, of these changes.
CNTY	As in RNDM but in this experiment all stations within the same country were given identical break points. This tests how well the system performs where coincident break points of the same magnitude exist at a number of neighboring stations.
META	As in RNDM but this time the system was run without metadata information. RNDM and META therefore provide a study of the extreme cases of complete metadata and no metadata, respectively.
SKEW	As in RNDM but in this experiment each break point was sampled from a normal distribution with mean of +0.5 K so that the break points were preferentially positive. This tests how well the system performs when the break points act to introduce a spurious trend in the station data.
SMALLSKEW	As in SKEW but the offset was −0.15. This removes the mean tropospheric temperature trend. The combination of SKEW and SMALLSKEW is also an important test that the system can recover trends that are both smaller or larger than the bias in the data.
MULTI	As in RNDM but in this experiment each station contained five break points separated by at least 2 yr.
MULTISKEW	The same as MULTI, but all the break points were positive and 95% of break points were less than 0.4 K in magnitude.
HATA	All adjustments applied to HadAT0 to create HadAT1 were applied inversely to the model test data. In this case break points exhibit a combination of the characteristics of the tests above, with the added complication that they are not vertically coherent. The record of metadata events used in HadAT was used here (i.e., the metadata was incomplete in comparison to the actual break points applied and also contained entries that were not associated with break points).

that the system is not simply achieving better statistics by removing all trends from the data. The small bias found in the adjustment estimates for these experiments (Table 2), coupled with remaining bias from the missed break points, means that some residual systematic trend error still remains following the homogenization. In MULTISKEW and MULTI the system performs poorly but does reduce the bias. These two experiments suggest that a major limitation on the recovery of trends from biased data is the breakpoint density rather than the average breakpoint magnitude. If the latter factor was more important, then we would expect MULTI, which consists of larger breaks, to have performed better than MULTISKEW, which consists of smaller breaks. The undetected break points are the greatest contribution to the remaining residual bias for MULTISKEW. Experiment HATA also has a relatively high breakpoint density, along with other complicating factors, and the trend error remains close to the unadjusted errors of 0.05 K decade^{-1} for global means, and 0.075 K decade^{-1} for tropical means.

TABLE 2. Summary statistics of nine model experiments: *P*(detection) is the probability of detecting a break point within one year of its actual occurrence; *P*(false) is the proportion of all break points identified that were false. Numbers in brackets refer to the total number of break points that meet the detection or false detection criteria. The rms difference and bias of the adjustment estimates are shown relative to the expected adjustment.

	Breakpoint detection (%)		Adjustment (K)	
Expt	*P*(detection)	*P*(false)	rms	Bias
UNADJ	0% (0)	100% (462)	0.50	0.01
RNDM	82% (779)	27% (281)	0.31	−0.02
CNTY	76% (725)	24% (227)	0.50	0.03
META	66% (629)	33% (308)	0.30	−0.01
SKEW	85% (809)	26% (284)	0.33	−0.1
SMALLSKEW	82% (781)	26% (276)	0.31	0.01
MULTI	61% (1455)	11% (171)	0.51	0.00
MULTISKEW	39% (926)	17% (195)	0.40	−0.06
HATA	58% (750)	34% (393)	0.36	−0.03

d. Summary

From the preceding analysis we can make a number of statements about the properties of the HadAT homogenization system and its limitations in recovering large-scale mean trends. The results presented in Table 2 are for all break points. If we consider only the tropical region (not shown) the results are similar, as evidenced by the trend errors shown in Fig. 4, so we are confident that these conclusions hold for both global and tropical means.

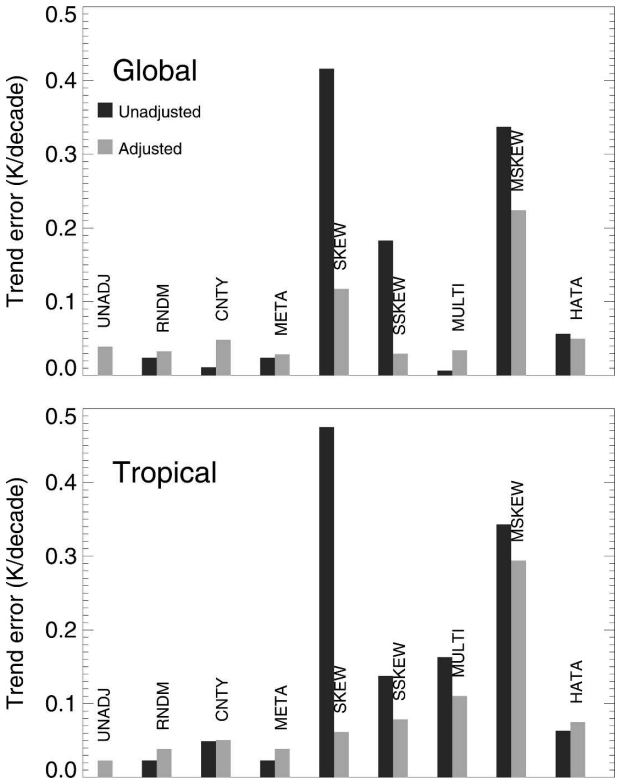

FIG. 4. Mean of nine model levels absolute trend error (K decade^{-1}) for (top) global and (bottom) tropical mean trends. The trend error is calculated as the difference of the model data with and without the addition of break points, both before and after the homogenization.

- Breakpoint detection is effective, particularly for breaks larger than 0.4 K. Good metadata will improve the detection of small (<0.4 K) break points and reduce false detection rates.
- Adjustment uncertainty of order 0.4 K means that small break points will not be adequately corrected. The presence of simultaneous break points within countries, or high breakpoint density, increases the adjustment uncertainty. The presence of a widespread systematic bias in break points can lead to a systematic underestimate of adjustments.
- In the absence of a systematic trend bias, the presence of adjustment uncertainty and false break points yields a trend uncertainty following homogenization of between 0.02 and 0.05 K decade^{-1}.
- When there is a systematic bias in trends, the homogenization will act to reduce the trend error. The extent to which the true underlying trend is recovered is strongly dependent upon the breakpoint density, and the number of undetected breaks that remain following homogenization, although other properties of the break points (magnitude, vertical profile, etc.) also play a role. When the homogenization system significantly alters the large-scale mean trends we would expect there to be some residual systematic bias, of the same sign as the shift produced but unknown magnitude.

6. Ensembles of random experiments

a. Model experiment

In the above analysis we have assessed the system capabilities and limitations under a single configuration of the available system parameters. To assess the sensitivity of results to different parameter settings we perform an ensemble of "random experiments" with system parameters randomly set to within reasonable bounds as defined in appendix A. In this way we can investigate the sensitivity of estimates of large-scale trends to changes in the methodology that will affect the number and type of break points detected. We conducted the first such ensemble on the SKEW experiment, described in Table 1, because its large trend bias provides scope for considerable spread in any such ensemble of homogenized trend estimates.

Fifty random experiments were performed to produce the ensemble results, and the spread of solutions for tropical means is shown in Fig. 5. The results are very similar for the global mean trends. The magnitude of the trend in the ensemble spans the space between the trend profiles for the original and biased model data. The shape of the vertical profile of trends for each of the 50 ensemble members closely resembles the shape of the original trend profile. The median of the ensemble does not adequately eliminate the systematic bias in the trends. This suggests that treating each ensemble member with equal weight will not fully account for systematic bias in trends. In this simple case there are parameter settings that can achieve an almost complete recovery of the original trend and alternative settings that have no significant impact on the biased data (many of which are likely to be conservative in the identification and/or adjustment of break points).

The analysis gives us some confidence that, in applying random ensembles to real data, we can gain useful information about the potential magnitude of uncertainty. Some homogenization system configurations will be more efficient in the removal of spurious trend bias. The challenge is therefore to establish a means of robustly distinguishing between the good and poor homogenizations.

b. Observations

A total of 200 random experiments were conducted on the IGRA merged (0000 and 1200 UTC) radiosonde

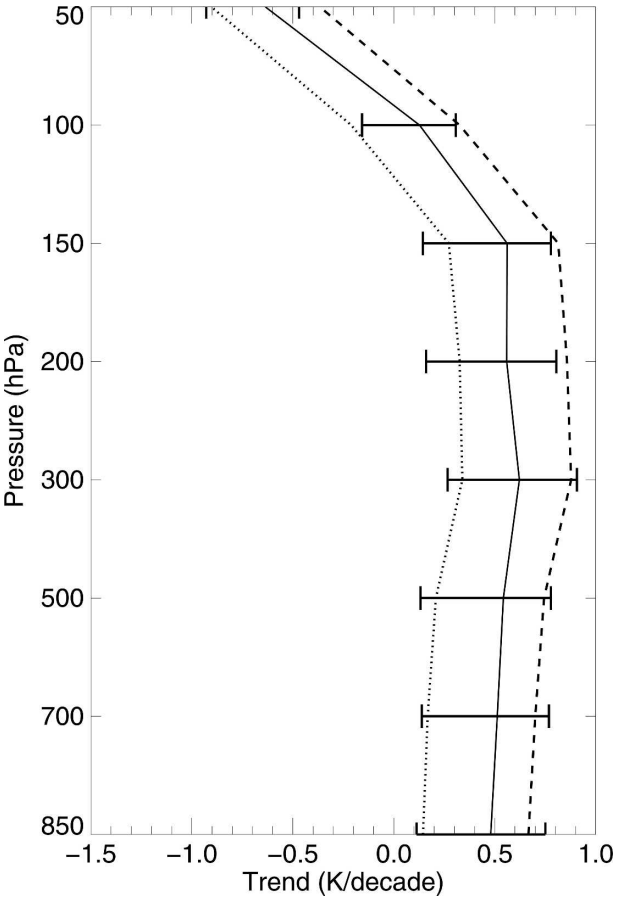

FIG. 5. Vertical profile of trends in tropical mean temperature for a set of 50 different versions of the automated homogenization applied to the SKEW experiment (see text). Dotted line is the original unadjusted model trend, dashed line is the biased model data, while the solid line represents the median and spread of 50 automated homogenizations.

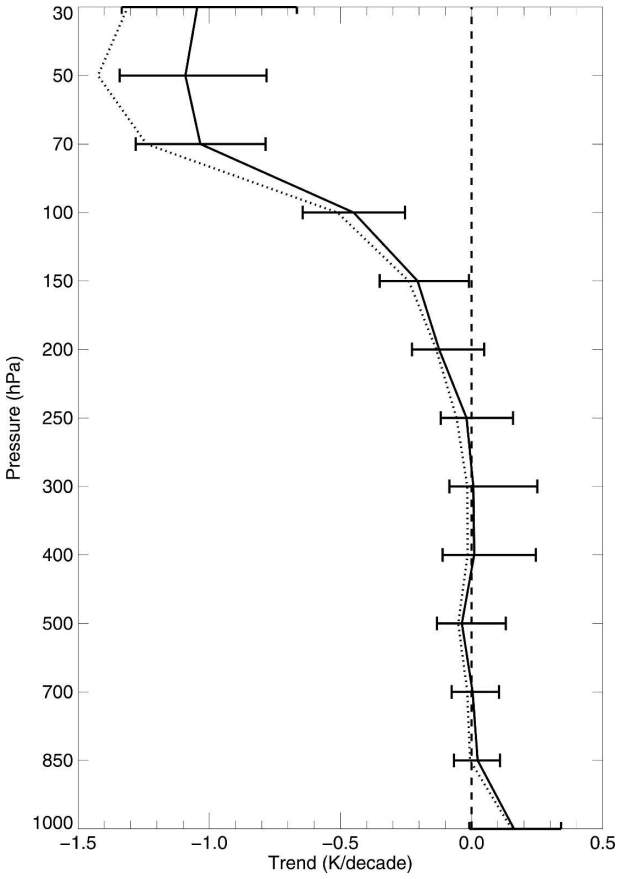

FIG. 6. Profile of trends for the tropics in the satellite era (1979–2003). The error bars represent the spread of 200 homogenizations; the thick solid line is the median of these 200 members; the dotted line is the unadjusted data.

dataset described in section 2. This provides a large sample with which to investigate system sensitivities. Profiles of the spread of trend estimates are presented in Fig. 6. The two-sigma spread of the ensembles are ± 0.03 K decade^{-1} in the global mean trends and ± 0.05 K decade^{-1} in the tropical mean trends for the lower troposphere for the satellite era. These are consistent with the random uncertainties derived from model experiments free from systematic bias (Fig. 4). These uncertainty estimates increase with height to ± 0.1 and ± 0.14 K decade^{-1} for the global and tropical stratosphere, respectively. They are also broadly consistent with the parametric uncertainty estimates presented for HadAT in Thorne et al. (2005a, Fig. 10). It should also be noted that, in the upper troposphere in particular, the ensemble of homogenized datasets show considerable skewness, with a greater tail toward warming values.

c. *Investigating systematic bias*

1) PARAMETER SENSITIVITIES

We know from the analysis so far that a neighbor-based homogenization as used here will struggle to recover true trends where systematic bias pervades the network used for the neighbor-based reference series (Fig. 4). We have also shown that particular system configurations achieve better recovery of trends in biased model data (Fig. 5) and result in increased tropospheric warming in the observations (Fig. 6). It is important therefore to objectively determine whether this occurs by chance or if particular parameter settings are beneficial. To summarize systematic trend differences resulting from parameter choices we computed trends from the IGRA homogenized data, during the satellite era, of MSU T2LT lower-troposphere temperatures (e.g., Mears et al. 2003) by weighting temperatures on the pressure levels. Static weighting functions were provided by the University of Alabama, Huntsville. Only 2

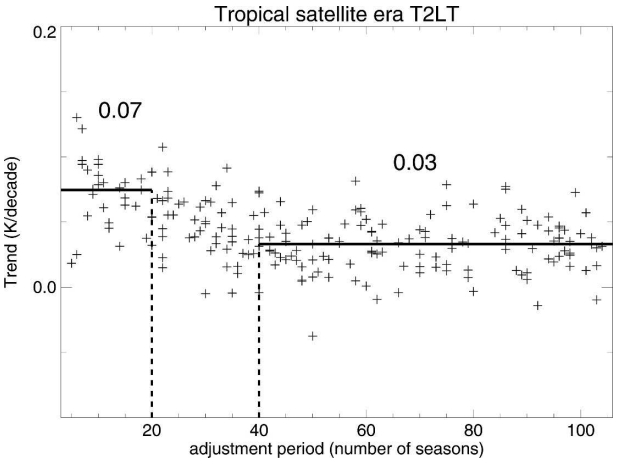

FIG. 7. Tropical T2LT trends for the satellite era as a function of the adjustment period set during each experiment. Also marked are the medians of the populations for which an adjustment period equal to or less than 20 seasons was set and for which an adjustment period equal to or greater than 40 seasons was set.

of the 14 tuneable parameters were found to have a discernable systematic impact on the large-scale trends, and both relate to the calculation of adjustments. Trends in the tropical lower troposphere are increased by 0.02 K decade^{-1} on average if we allow the system to recalculate all adjustments at every iteration, rather than applying them only on the first iteration they are found (see adjustment method in Table A1). Given that the neighbor reference series is expected to improve with each iteration we expect such an adaptive method to be preferable. For the model experiments in section 6a the residual trend error following homogenization is 0.08 K decade^{-1} using an adaptive method and 0.43 K decade^{-1} using a nonadaptive method (the trend error in the unadjusted data is 0.53 K decade^{-1}).

The second parameter is the adjustment period, which is the time period used to estimate the adjustment factor. Estimated tropical mean trends from observations in the lower troposphere were increased by 0.04 K decade^{-1} with an adjustment period of less than 20 seasons (5 yr) compared with periods greater than 40 seasons (10 yr) (Fig. 7). While longer periods should reduce the noise in adjustment estimates, if systematic biases exist in the network these may be aliased into the neighbor composites and potentially make longer adjustment periods undesirable. A clear distinction in trend recovery for different adjustment periods was not apparent in the model ensemble.

2) SEPARATING DAY AND NIGHT

To investigate further the potential magnitude of systematic bias in the trend estimates we separated the day and night data, which are known to be biased in relation to one another (e.g., Sherwood et al. 2005; Randel and Wu 2006). In Fig. 8 we show time series of global mean temperature anomalies for both day and night. In this example we have conducted the homogenization using the same system configuration as used in sections 4 and 5 (i.e., the HadAT-like parameter set). The trend of day relative to night in the unadjusted data changes sign coincident with the inception of the satellite record. It is clearly not true that unadjusted day data have a simple warm bias that decreases over time. The biases are significantly larger in the raw data than the adjusted data, which implies that the homogenization has successfully removed many of the inconsistencies, even when the day and night are adjusted independently. This is an encouraging result, supporting our earlier conclusions that the homogenization will at least remove part of the systematic bias and that HadAT will have at least partially remedied day/night biases, even without explicitly considering them. In an additional experiment we used the adjusted night data as the neighbor reference field for the day time dataset (bottom panel in Fig. 8). In this case we found that the day–night trend discrepancy is removed, suggesting that in the presence of a fixed reference network free from systematic bias we can expect our system to effectively remove such bias from the observations. However, we caution that the night data are not necessarily a suitable transfer standard for the day data and may also contain systematic inhomogeneities. For example, in recent decades temperature probes have changed from being painted white to being metallic. This affects the IR absorption characteristics of the sensors, which can result in significant bias in both day and nighttime measurements (J. Nash 2006, personal communication).

7. Spanning the range and comparison to other datasets

In the previous sections we have identified limitations of the homogenization system and some key potential sources of systematic bias. We now use this information to create an ensemble of homogenizations intended to span the possible trend solutions, using the automated system, highlighting day–night bias, and optimal system parameter settings. This should provide a plausible uncertainty estimate for HadAT-like radiosonde climate records. Two sets of 50 experiments were conducted on each for day and night. We used the same 50 experiments applied to the model data in section 6a, but fixed the adjustment method and adjustment period (section 6c). First the 50 experiments were set to have adaptive adjustments with a short adjustment period

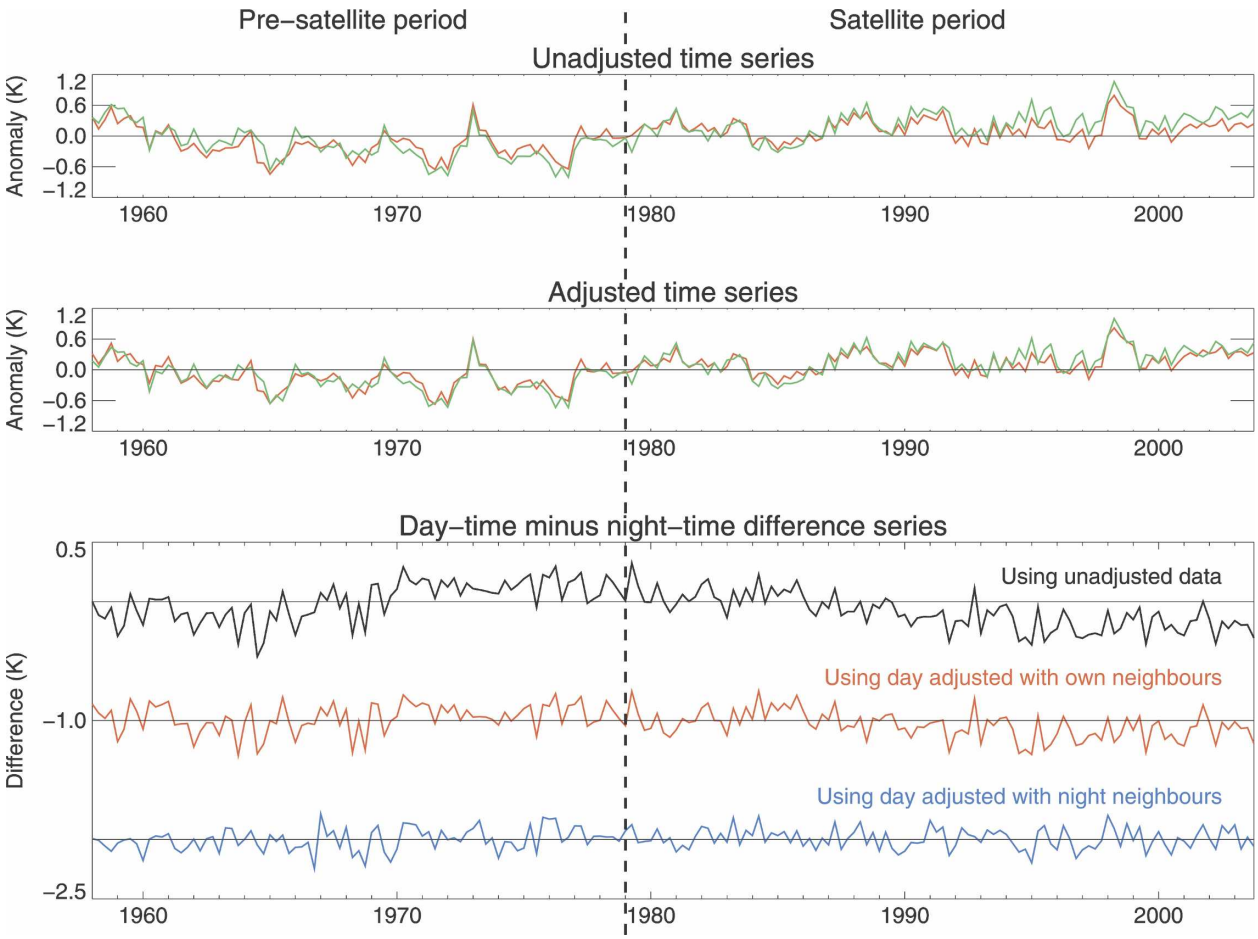

FIG. 8. Time series of global mean temperature at 500 hPa: (top) The unadjusted daytime (red) and nighttime (green) data; (middle) the same data after adjustment using their own neighbors; (bottom) differences between the unadjusted daytime and nighttime series (upper), the adjusted daytime (using own neighbors) and nighttime series (middle; offset by −1 K), and the adjusted daytime (using nighttime neighbors) and nighttime series (bottom; offset by −2 K). The vertical dashed black line denotes the beginning of the satellite period (1979).

(<20 seasons), called the "max" ensemble as this was expected to produce the greatest tropical warming. The same 50 experiments were then run with nonadaptive adjustments and a long adjustment period (>40 seasons), called the "min" ensemble.

We compare our day and night ensembles with a range of other estimates. Three of these come from the satellite MSU instruments: University of Alabama, Huntsville (UAH) version 5.2 (Christy and Norris 2004, 2006), Remote Sensing Systems (RSS), version 2.1 (Mears et al. 2003; Mears and Wentz 2005), and the University of Maryland (UMd; Vinnikov et al. 2006). Three estimates come from radiosonde datasets: the Radiosonde Atmospheric Temperature Products for Assessing Climate (RATPAC; Lanzante et al. 2003; Free et al. 2005), Radiosonde Observation Correction using Reanalyses (RAOBCORE), version 1.2 (Haimberger 2007), and HadAT1 (Thorne et al. 2005a). These are all blends of day and nighttime observations. For comparison the radiosonde datasets have been vertically averaged with weightings to create equivalent MSU retrieved bulk temperatures for T4 (stratosphere), T2 (troposphere), and T2LT (lower troposphere). For reference we also show the trend from the Hadley Centre Climatic Research Unit, version 3 (HadCRUT3), surface records (Brohan et al. 2006), and theoretical expectation from an ensemble of climate models (Santer et al. 2005).

Figure 9 shows the spread of trend estimates from this ensemble of homogenizations for the tropical satellite era. Differences between the max and min ensembles grow with height, as would be expected if the daytime radiation biases are a major contributing factor. The daytime radiosonde estimates are biased low for T4 (the stratosphere) relative to MSU instrument estimates. Mears et al. (2006) concluded that this most

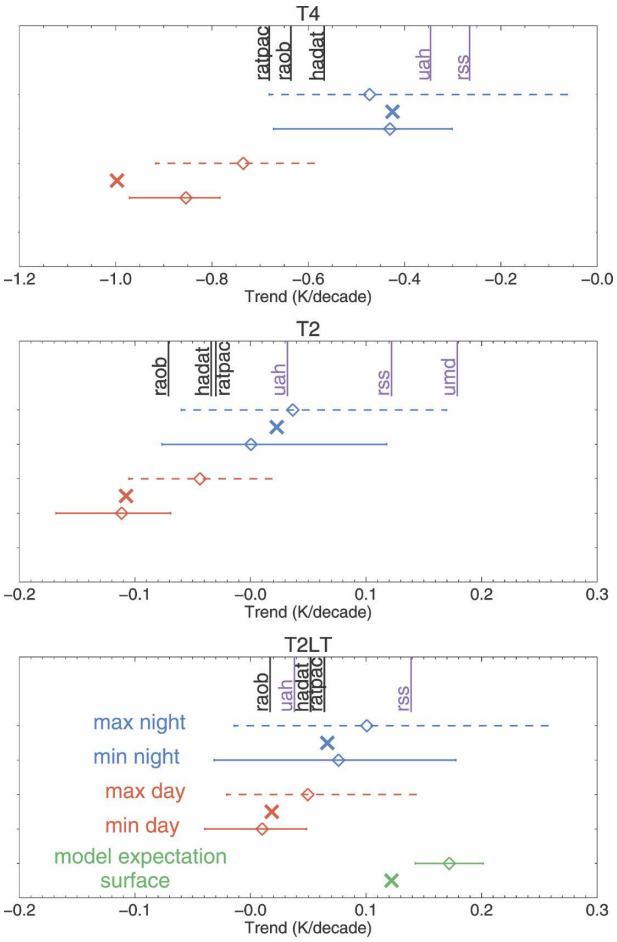

FIG. 9. The spread of trend estimates for the tropics in the satellite era (1979–2003) for the maximum ensemble (dashed line) and minimum ensemble (solid line), for day (red) and night (blue). Median estimates are given by diamonds, and the uncorrected trends are shown as crosses. Trends are presented for MSU equivalent bulk temperatures from (top) T4, peaking in the lower stratosphere, T2 peaking in the mid upper troposphere, and T2LT peaking in the low troposphere. For each channel other MSU (purple) and radiosonde (black) estimates (see text) are also presented. T2LT includes an additional green cross denoting the HadCRUT3 surface record trends, and a range of theoretical estimates of what climate models predict for the T2LT trend given the HadCRUT3 surface warming. Please note the change of scale for T4.

likely relates to pervasive cooling biases in the radiosondes within the stratosphere. The maximum ensemble clearly shifts the trend in raw observations toward closer agreement with the satellite estimates. However, assuming the MSU measures are grossly adequate, it does not move them far enough. This effect also impacts T2, which has about 10%–15% of its weighting from the stratosphere. The day–night trend discrepancy is reduced in nearly all experiments for T4.

For the T2LT trends (Fig. 9) there is a distinction between the max and min ensembles for day, with the median max ensemble resulting in a T2LT trend 0.05 K decade^{-1} larger than the median of the min. The day–night trend difference is reduced in 50% of the max ensemble members, but for only 28% of the min ensemble members. This is further evidence that the max ensemble is a more robust homogenization method. If we then consider only the max experiments in which the day–night trend discrepancy is reduced, the median daytime T2LT trend estimate (0.07 K decade^{-1}) is 0.02 K decade^{-1} larger than the median of all the daytime max experiments (0.05 K decade^{-1}). There is no preferential shift in the night-only data but the uncertainties are larger due to the reduced spatial sampling of the night data.

Our analysis of the system suggests that where systematic bias is prevalent, the system is likely to only partly recover the underlying natural trend (Figs. 4 and 5), and there is currently no a priori way of estimating by how much. Figure 5 also showed that the median of an ensemble of trend estimates from different homogenizations is unlikely to be a reliable indicator of the magnitude of systematic bias, and in this analysis we have used the median trends simply to quantify the impact of particular system configurations. Therefore, the results presented in Fig. 9 suggest that homogenized radiosonde trends from HadAT are potentially underestimated due to daytime biases, and the spread of trend estimates for day and night suggest that a combination of the homogenization uncertainty and residual systematic bias is sufficient to encapsulate much of the discrepancy between HadAT, theoretical expectation, and surface temperature trends. Further analysis is required to attempt to reduce this uncertainty.

8. Discussion and conclusions

We have presented an automated method for the identification and adjustment of break points in radiosonde temperature data, which we propose can be used to investigate uncertainty in estimates of multidecadal trends in historical temperature records. The process implemented here is based closely on that used in the generation of HadAT (Thorne et al. 2005a) in order to provide a benchmark for assessing the system.

The principal advantages of this approach are its flexibility and reproducibility, and the short time taken to produce a dataset (a few hours rather than a few years). It can be run under many different configurations in order to test the sensitivity of diagnostics such as linear trend estimates, to the process of homogenization. Our aim has not been to provide a single "best guess" assessment of the evolution of the historical temperature

record but rather to provide a method that we can use to objectively assess the homogenization of HadAT.

With appropriate tuning of the system we have shown that we are able to replicate many of the decisions employed in the homogenization of HadAT, with a detection rate of approximately 60% and a false detection rate of 14%. This gives us confidence that the automated system can be used to more rigorously assess limitations and uncertainties in the homogenization of HadAT-like datasets.

Break points can be successfully identified and adjusted by the system where they are at least 0.4 K in magnitude, but are poorly resolved if much smaller than this. Good metadata records will aid in the identification of these smaller break points (see also Free et al. 2005; Haimberger 2007) and reduce the false detection rates. Coincident break points at neighboring stations or high breakpoint density is not hugely detrimental to the detection of break points, but will increase the uncertainty on adjustment estimates. In the absence of systematic bias, trend uncertainty from the homogenization process is 0.02–0.05 K decade^{-1}. In the presence of a pervasive systematic bias the homogenization is likely to underestimate adjustments and will reduce, but not remove, any bias in large-scale mean trends. The density of break points and consequently the number of undetected break points following homogenization is a major limiting factor for the recovery of trends in such instances.

A number of factors discussed in this work are general to all related attempts to homogenize radiosonde temperature records:

- Biases exist in day relative to night radiosonde data (e.g., Sherwood et al. 2005; Randel and Wu 2006; Lanzante et al. 2003; Haimberger 2007).
- Quality of the background or neighbor reference is a major influence on trend recovery (e.g., Thorne et al. 2005a; Haimberger 2007).
- The time interval used for estimating the breakpoint adjustment is an important consideration for homogenization methods (e.g., Haimberger 2007).
- Methodological choices can result in significant parametric uncertainty in radiosonde trend estimates, and methods should, if possible, be objectively tested in their ability to recover climate signals from data with trend biases. We have done this for the HadAT dataset.

Further development of our automated homogenization method could be achieved by considering ways to improve the reliability of the background reference by accounting for transient regional climate anomalies, or explicitly reducing the influence of break points in the neighbor reference series. We should also consider additional independent evidence (either physical or statistical) that might be placed on the estimation of break points. One example may be the thermal wind relation (Allen and Sherwood 2007).

Our analysis provides further evidence that a combination of systematic and random uncertainties relating to the removal of biases using a HadAT-like methodology are sufficiently large to explain the tropical trend discrepancy between HadAT and estimates from other observational platforms, theoretical expectations, and trends at the surface. A previous assessment of trends and uncertainty in HadAT (Thorne et al. 2005a) makes a good estimate of the random homogenization uncertainty but, due to limitations of the homogenization method, is likely to have underestimated the daytime systematic bias component and, therefore, the resultant trends in the tropical troposphere. Further analysis of the homogenization system is currently under way with the aim of determining objectively the optimum configurations for robustly recovering large-scale mean trends in radiosonde temperature records and more appropriate ways to reject unreasonable homogenizations from an ensemble of the automated system.

Acknowledgments. Thanks to Steven Sherwood and William Ingram for comments on an earlier draft, and four anonymous reviewers. The work of Mark McCarthy, David Parker, Simon Tett, and Holly Titchner was funded by the U.K. Department of the Environment, Food, and Rural Affairs (Defra) and Ministry of Defense (MoD) under the joint Defra and MoD integrated climate program, (Defra) GA01101, (MoD) CBC/2B/0417_Annex_C5. This paper is British Crown copyright. Leo Haimberger was funded by Contract P18120-N10 of the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (FWF)

APPENDIX A

Homogenization System Parameters

The automated homogenization system contains a number of sensitivities and subjective parameters that affect various components of the homogenization process. These sensitivities are outlined in Table A1, along with a description of how the various parameters were set for this study.

APPENDIX B

Assessing Bias Adjustment Estimates

After the statistical breakpoint detection and estimation of adjustments, a number of quality assessments are conducted on the adjustment estimates to check that they arise from the station data, not the neighbor composite, and are not greatly affected by outliers in

TABLE A1. A summary of key parameters in the automated homogenization system.

Parameter	Default setting	Alternative settings	Description
Neighbor weighting coefficients	Derived from NCEP–NCAR reanalysis fields	Derived from ERA-40 fields	Weighting coefficients for possible neighbor stations.
Country	Off	On	Excludes any neighbor stations within the same country as the target station.
Metadata	Off	On	Excludes any neighbor stations with similar metadata records as the target station.
K–S window width	15	8–20	Number of seasons used for the K–S test used to assign break points.
Metadata weighting	0.4	0–1	Weighting given to metadata events during the breakpoint identification procedure (0: no weight, 1: break point at every metadata event).
Metadata_function	Gaussian	Exponential or step	Shape of inversion in the metadata statistic series at known events.
Vary metadata background	Off	On	Alters the background value of the metadata probability series for each station based on the number of metadata events (i.e., penalizes stations with poor metadata records).
Range	8	6–20	Minimum number of seasons required between each break point.
Critical value	0.01	0.001–0.100	Initial critical threshold used to identify break points in the first iteration.
Max iteration	5	1–15	Number of iterations performed.
Iteration step	0.03	0.001–0.050	Increment that the critical value is increased by with each iteration.
Adjustment method	Nonadaptive	Semiadaptive or adaptive	Adjustment method used. Adaptive recalculates all adjustments at each iteration. Semiadaptive recalculates only if the break point is found again at a later iteration. Nonadaptive calculates adjustment only when the break point is first found.
Adjustment_period	40	5–105	Number of seasons either side of each break point used to calculate an adjustment factor.
Adjustment_threshold	[5, 8]	[0, 0]–[7, 11]	Thresholds for determining whether an adjustment should be applied or not based on a points scoring system (appendix B).

the data. A modified bootstrap method is used to make 200 estimates of the adjustment by dropping out a random amount of randomly distributed data points either side of the break point (Thorne et al. 2005a). Eleven tests were carried out on this population of adjustment estimates, each with a positive or negative outcome. The number of positive results was counted for each pressure level and then averaged over all levels to provide a breakpoint score. In order for an adjustment to be applied to the data this score had to be larger than a predefined value (default of 5, see appendix A). Alternatively if any single level scored highly (default of 8) then the break point was also applied. The first two tests were employed directly in HadAT and the others were designed to replicate the key points of subjective evidence used in the manual process of accepting or rejecting break points that had been identified by the statistical detection and adjustment algorithm.

- Is the adjustment significantly different from zero? (1 point)
- Is the population of adjustment estimates approximately normally distributed (3 points)
  - Is the difference between the mean and median less than 25% of the magnitude of the median?
  - Are the 5th and 95th percentiles equidistant, to within 25%, from the median?
  - Are the first and 99th percentiles between 1.1 and 1.9 times the distance of the fifth and 95th from the median?
- Is the break point an artifact of the station series, or does it originate from errors in the neighbor composite? (5 points)
  - Is the K–S statistic from the station minus neighbor difference series smaller than the K–S statistic calculated using the neighbor composite series only?

- Is the K–S test statistic calculated from the stations series smaller than that calculated from the neighbor composite series?
- Is the adjustment estimate calculated from the difference series greater than an equivalent estimate calculated from the neighbor composite series?
- Is the adjustment estimate calculated from the station series greater than that calculated from the neighbor composite series?
- Is the adjustment estimate calculated from the neighbor composite series less than 25% of the magnitude of the estimated adjustment from the difference series (i.e., close to zero).

- Is the break point associated with a metadata event? (1 point)
- Is the adjustment vertically coherent, that is, within 25% of the adjustment estimated from linear interpolation of neighboring pressure levels)? (1 point)

REFERENCES

Allen, R. J., and S. C. Sherwood, 2007: Utility of radiosonde wind data in representing climatological variations of tropospheric temperature and barclinicity in the western tropical Pacific. *J. Climate,* **20,** 5229–5243.

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.,* **111,** D12106, doi:10.1029/2005JD006548.

Christy, J. R., and W. B. Norris, 2004: What may we conclude about global tropospheric temperature trends? *Geophys. Res. Lett.,* **31,** L06211, doi:10.1029/2003GL019361.

——, and ——, 2006: Satellite and VIZ–radiosonde intercomparisons for diagnosis of nonclimatic influences. *J. Atmos. Oceanic Technol.,* **23,** 1181–1194.

Durre, I., R. S. Vose, and D. B. Wuertz, 2006: Overview of the integrated global radiosonde archive. *J. Climate,* **19,** 53–68.

Free, M., and D. J. Seidel, 2005: Causes of differing temperature trends in radiosonde upper air datasets. *J. Geophys. Res.,* **110,** D07101, doi:10.1029/2004JD005481.

——, ——, J. K. Angell, J. Lanzante, I. Durre, and T. C. Peterson, 2005: Radiosonde Atmospheric Temperature Products for Assessing Climate (RATPAC): A new data set of large-area anomaly time series. *J. Geophys. Res.,* **110,** D22101, doi:10.1029/2005JD006169.

Gaffen, D. J., M. A. Sargent, R. E. Habermann, and J. R. Lanzante, 2000: Sensitivity of tropospheric and stratospheric temperature trends to radiosonde data quality. *J. Climate,* **13,** 1776–1796.

Haimberger, L., 2007: Homogenization of radiosonde temperature time series using innovation statistics. *J. Climate,* **20,** 1377–1403.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

Karl, T. R., S. J. Hassol, C. D. Miller, and W. L. Murray, Eds., 2006: *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences.* Climate Change Science Program and the Subcommittee on Global Change Research, 164 pp.

Lanzante, J. R., 1996: Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *Int. J. Climatol.,* **16,** 1197–1226.

——, S. A. Klein, and D. J. Seidel, 2003: Temporal homogenization of monthly radiosonde temperature data. Part I: Methodology. *J. Climate,* **16,** 224–240.

Mears, C. A., and F. J. Wentz, 2005: The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science,* **309,** 1548–1551.

——, M. C. Schabel, and F. J. Wentz, 2003: A reanalysis of the MSU channel 2 tropospheric temperature record. *J. Climate,* **16,** 3650–3664.

——, C. E. Forest, R. W. Spencer, R. S. Vose, and R. W. Reynolds, 2006: What is our understanding of the contribution made by observational or methodological uncertainties to the previously reported vertical differences in temperature trends? *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences,* T. R. Karl et al., Eds., Climate Change Science Program and the Subcommittee on Global Change Research, 71–88.

Parker, D. E., M. Gordon, D. P. N. Cullum, D. M. H. Sexton, C. K. Folland, and N. Rayner, 1997: A new global gridded radiosonde temperature data base and recent temperature trends. *Geophys. Res. Lett.,* **24,** 1499–1502.

Pope, V. D., M. L. Gallani, P. R. Rowntree, and R. A. Stratton, 2000: The impact of new physical parametrizations in the Hadley Centre climate model—HadAM3. *Climate Dyn.,* **16,** 123–146.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in FORTRAN: The Art of Scientific Computing.* 2nd ed. Cambridge University Press, 963 pp.

Randel, W. J., and F. Wu, 2006: Biases in stratospheric and tropospheric temperature trends derived from historical radiosonde data. *J. Climate,* **19,** 2094–2104.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.,* **108,** 4407, doi:10.1029/2002JD002670.

Santer, B. D., and Coauthors, 2005: Amplification of surface temperature trends and variability in the tropical atmosphere. *Science,* **309,** 1551–1556.

Sherwood, S. C., J. R. Lanzante, and C. L. Meyer, 2005: Radiosonde daytime biases and late-20th century warming. *Science,* **309,** 1556–1559.

Tett, S. F. B., and Coauthors, 2007: The impact of natural and anthropogenic forcings on climate and hydrology since 1550. *Climate Dyn.,* **28,** 3–34.

Thorne, P. W., D. E. Parker, S. F. B. Tett, P. D. Jones, M. McCarthy, H. Coleman, and P. Brohan, 2005a: Revisiting radiosonde upper-air temperatures from 1958 to 2002. *J. Geophys. Res.,* **110,** D18105, doi:10.1029/2004JD005753.

——, ——, J. R. Christy, and C. A. Mears, 2005b: Uncertainties in climate trends: Lessons from upper-air temperature records. *Bull. Amer. Meteor. Soc.,* **86,** 1437–1442.

Uppala, S. M., and Coauthors, 2005: The ERA-40 re-analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012.

Vinnikov, K. Y., N. C. Grody, A. Robock, R. J. Stouffer, P. D. Jones, and M. D. Goldberg, 2006: Temperature trends at the surface and in the troposphere. *J. Geophys. Res.,* **111,** D03106, doi:10.1029/2005JD006392.